Основы искусственного интеллекта

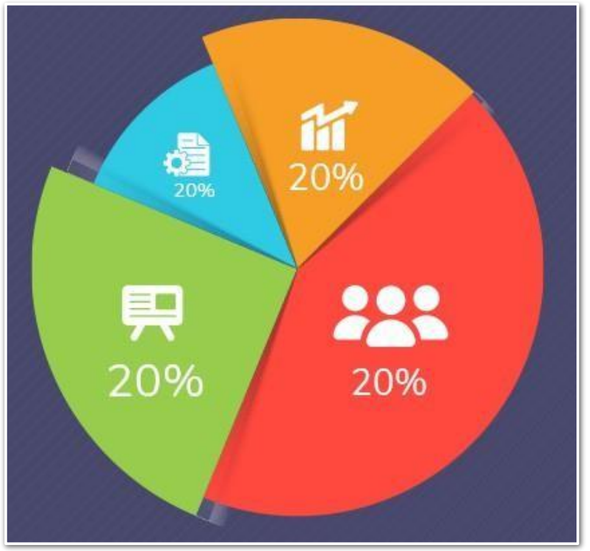
Основы инференциальной статистики

Введение

Что такое инференциальная статистика?

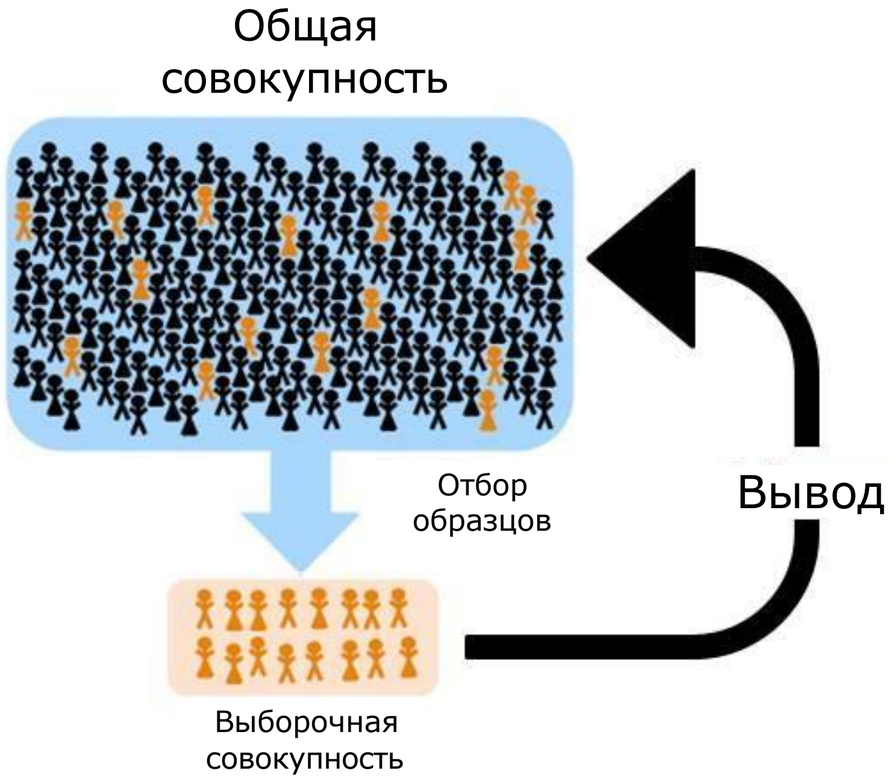
- **Инференциальная статистика** (statistical inference) позволяет делать выводы об общей совокупности на основе выборки.

- В отличие от описательной статистики, **инференицальная статистика** позволяет нам делать прогнозы, делать выводы и обобщать выводы, полученные на основе выборки, на всю совокупность.



Для чего нужна инференциальная статистика?

- Представьте, что вы анализируете поведение клиентов для платформы электронной коммерции.
- Непрактично собирать данные от каждого отдельного клиента **(общей совокупности).**
- Вместо этого вы собираете данные от небольшой группы клиентов **(выборочной совокупности).**
- Но как вы можете уверенно делать заявления о всей общей совокупности, основываясь только на маленькой выборке?
- Вот где в игру вступает инференциальная статистика.



Для чего нужна инференциальная статистика?

- **Инференциальная статистика помогает нам:**
  - **Обобщать выводы** из выборки на совокупность.
  - **Проверять гипотезы** для подтверждения предположений или утверждений о данных.
  - **Количественно оценивать неопределенность**, вычисляя доверительные интервалы и p-значения.
  - **Делать прогнозы** с использованием статистических моделей.

Для чего нужна инференциальная статистика?

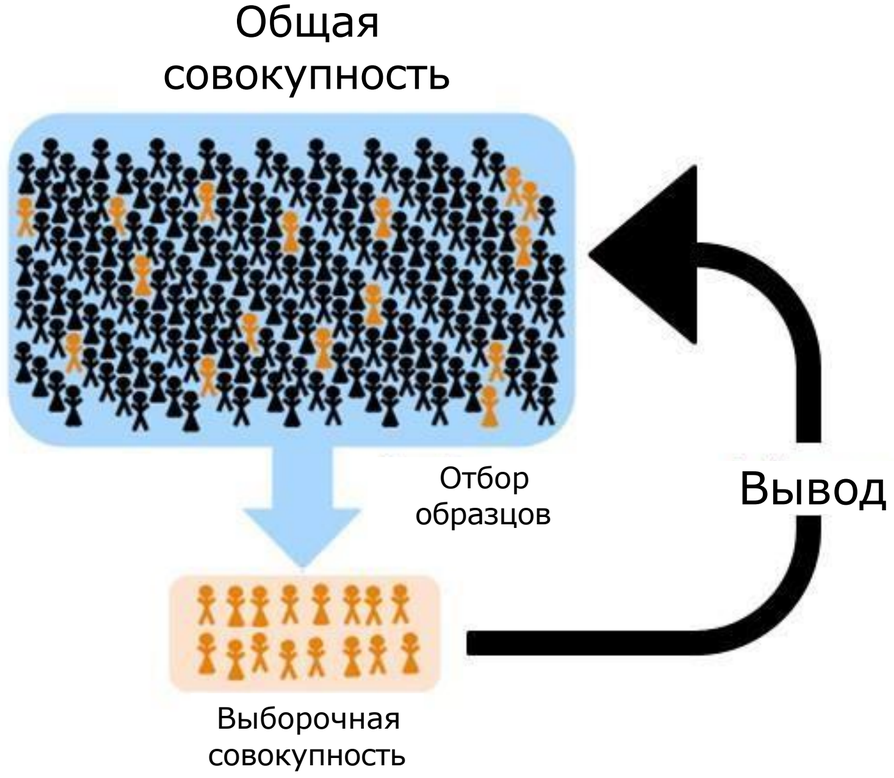
- **Инференциальная статистика** предоставляет инструменты, позволяющие выйти за рамки имеющихся у вас данных и принимать обоснованные решения относительно данных, которых у вас нет.

Основные концепции

- **Общая совокупность** (population) — вся группа объектов, о которых делаем вывод.
- **Выборка** (sample) — подмножество генеральной совокупности.
- **Параметры** — характеристики общей совокупности (например, среднее, дисперсия).
- **Статистики** — характеристики выборки, используемые для оценки параметров.

Выборка и распределение выборки

- **Методы выборки** наряду с **распределениями выборки** обеспечивают контекст для того, как собираются образцы, прежде чем обсуждать, как ведут себя их статистики (распределения выборки).
- Связь теории с реальными процессами сбора данных (методы выборки).



Что такое выборка?

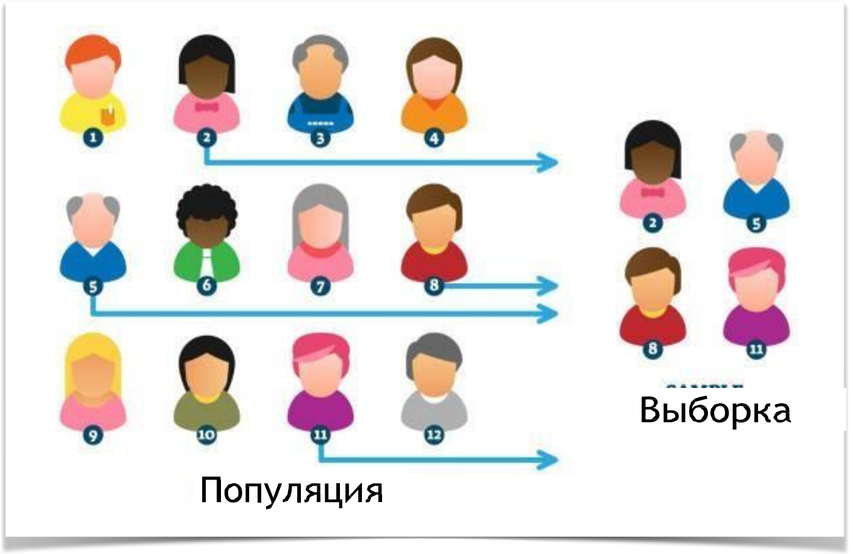
- Он включает в себя выбор подмножества индивидуумов или точек данных из более крупной популяции для изучения.
- Поскольку анализ всей популяции часто нецелесообразен или невозможен, выборка позволяет делать выводы о популяции на основе этой меньшей подгруппы.
- **Репрезентативная выборка** минимизирует смещение и гарантирует, что выводы, полученные из выборки, могут быть обобщены на всю популяцию.

Типы методов выборки

- Чтобы выбрать репрезентативную выборку, мы используем различные **методы выборки**, которые можно в целом разделить на две категории:

  1. **вероятностная выборка**

  2. **невероятностная выборка**



- При вероятностной выборке **каждый индивидуум из общей совокупности имеет известный и равный шанс быть выбранным.** Это гарантирует, что выборка будет беспристрастной и репрезентативной

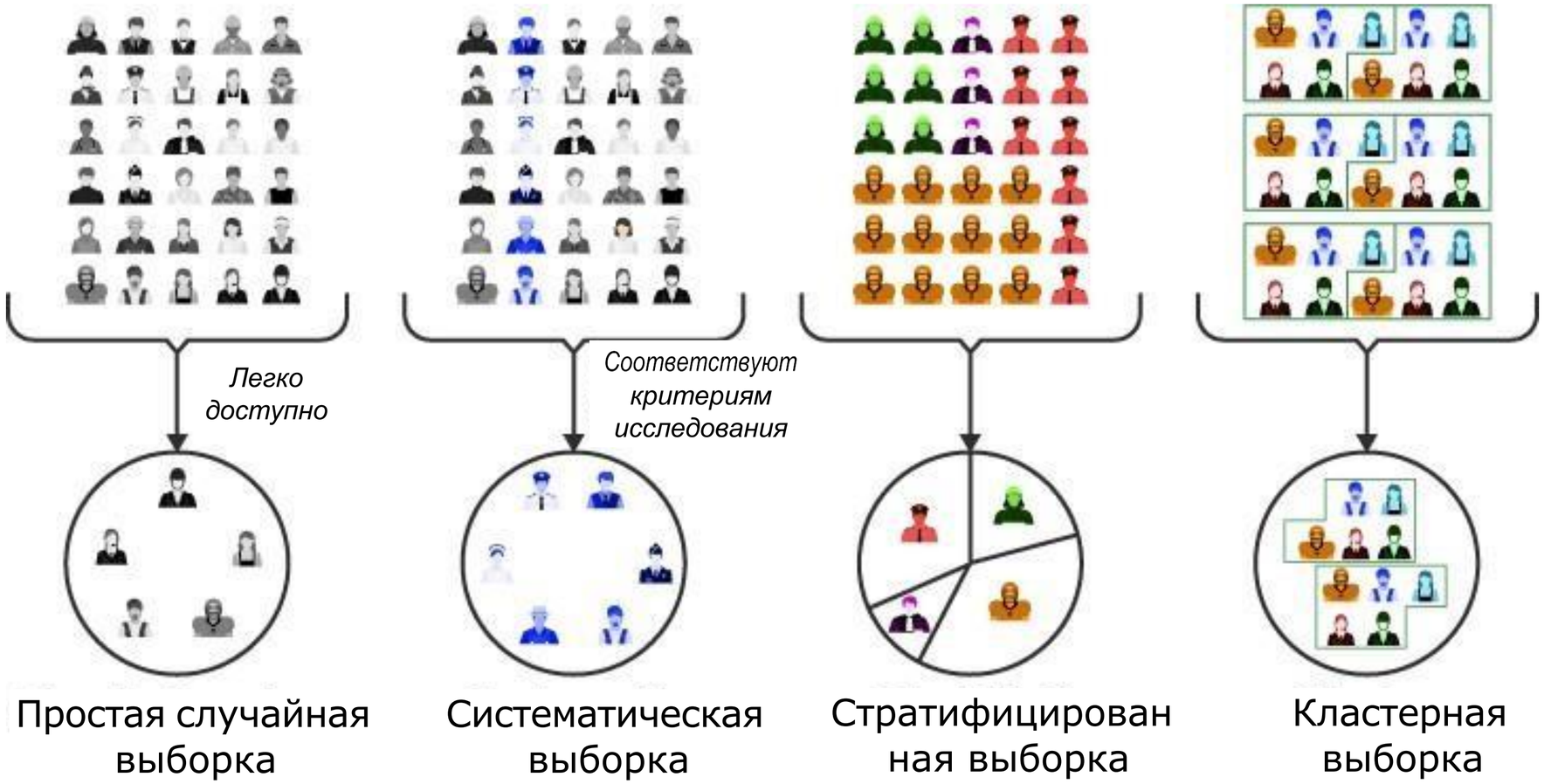
Типы методов выборки

Вероятностная выборка

Метод выборки	Описание	Пример
Простая случайная выборка	Каждый элемент имеет равную вероятность быть выбранным.	Случайный выбор 100 клиентов из базы данных с помощью генератора случайных чисел
Стратифицированная выборка	Население делится на подгруппы (страты) на основе определенных характеристик, и выборки берутся пропорционально из каждой подгруппы.	Разделение клиентов на возрастные группы (например, 18–25, 26–40) и выборка пропорционально из каждой группы.
Систематическая выборка	Каждый n-й человек выбирается из списка после выбора случайной начальной точки.	Выбор каждого 10-го клиента из базы данных.
Кластерная выборка	Население делится на кластеры (например, географические регионы), и целые кластеры выбираются случайным образом для анализа.	Выбор двух городов случайным образом и опрос всех клиентов в этих городах.

Типы методов выборки

Вероятностная выборка



Типы методов выборки

Невероятностная выборка

- При невероятностной выборке не все люди имеют равные шансы быть выбранными.
- Эти методы часто быстрее и дешевле, но могут вносить смещение.

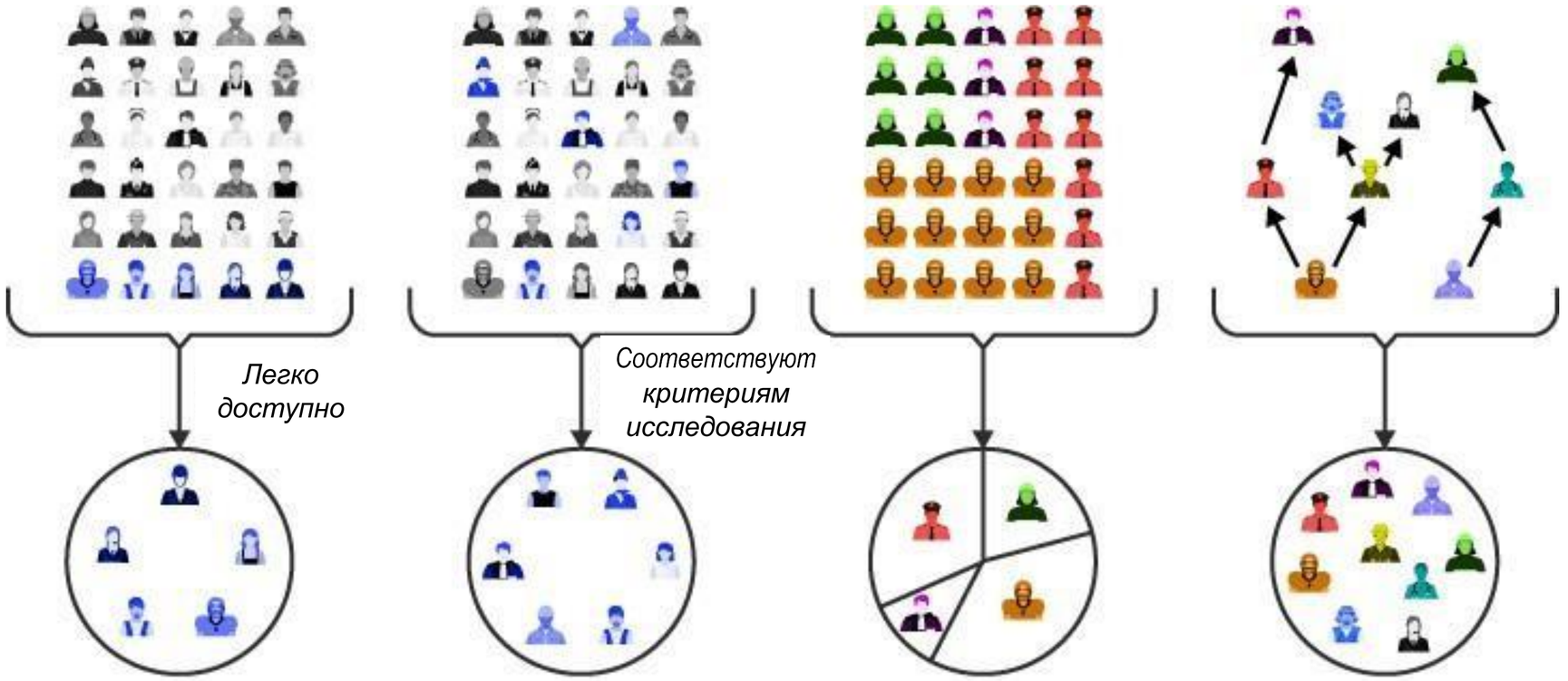
Типы методов выборки

Невероятностная выборка

Метод выборки	Описание	Пример
Выборка по удобству	Выбор элементов, к которым проще всего получить доступ.	Опрос людей в торговом центре или использование легкодоступных данных.
Оценочная (целевая) выборка	Исследователь выбирает участников на основе своего суждения о том, кто может предоставить полезную информацию.	Интервьюирование отраслевых экспертов для понимания тенденций рынка.
Выборка по квотам	Гарантирует, что в выборке представлены определенные подгруппы, но выбор внутри этих подгрупп не является случайным.	Опрос 50 мужчин и 50 женщин без случайного отбора внутри каждой группы.
Выборка по принципу снежного кома	Участники привлекают других участников, часто используется для труднодоступных групп.	Опрос членов нишевых сообществ или социальных сетей

Типы методов выборки

Невероятностная выборка



Выборка по удобству

Оценочная (целевая) выборка

Выборка по квотам

Выборка по принципу снежного кома

Методы в инференциальной статистике

- Помогает нам ***делать прогнозы, проверять предположения и количественно оценивать неопределенность при работе с данными, взятыми из более крупной популяции.***

- Эти инструменты необходимы для таких задач, как:

  - ***оценка модели***

  - ***сравнение производительности и принятие решений в рабочих процессах машинного обучения***

Методы в инференциальной статистике

Доверительные интервалы

- **Доверительные интервалы:** предоставляют диапазон значений, в который истинный параметр популяции, вероятно, попадет с определенным уровнем уверенности.

- Например, 95% доверительный интервал для среднего значения популяции указывает, что мы на 95% уверены, что истинное среднее значение популяции попадает в этот интервал.

- Доверительные интервалы используются для количественной оценки неопределенности прогнозов модели или оценок параметров.

- Например, при оценке производительности модели (точности или частоты ошибок) доверительные интервалы помогают нам понять надежность метрик.

Доверительные интервалы

Пример

- **Вопрос:** Насколько сильна корреляция между ростом (в дюймах) и весом (в фунтах) у американских подростков?

- Две интересующие нас переменные: (1) **рост в дюймах** и (2) **вес в фунтах.** Обе являются количественными переменными. Интересующий нас параметр — это корреляция между этими двумя переменными.

- Нам не дается конкретная корреляция для проверки. Нас просят оценить силу корреляции. Соответствующая процедура здесь — доверительный интервал для корреляции.

Методы в инференциальной статистике

Проверка гипотез

- **Проверка гипотез** — это статистический метод, используемый для проверки предположений или утверждений о параметре популяции. Он широко используется для оценки значимости результатов.
- Например:
  - **Проверка того, значительно ли новая функция улучшает производительность модели.**
  - **Сравнение точности двух моделей для определения того, является ли одна из них статистически лучше другой.**
  - **Проверка предположений о распределении данных (например, нормальности) перед применением определенных алгоритмов.**
  - **Она включает формулирование гипотез, использование статистических тестов и вычисление P-значений**

Проверка Гипотез

Нулевая и Альтернативная гипотеза

- **Нулевая гипотеза:** предположение по умолчанию (например, «Между двумя моделями нет разницы»).
- **Альтернативная гипотеза:** утверждение, которое вы хотите проверить (например, «Модель А работает лучше, чем модель В»).

Проверка Гипотез

Нулевая и Альтернативная гипотеза

- Статистические тесты, такие как z-тесты, t-тесты или тесты чи-квадрат, используются для определения того, достаточно ли доказательств для отклонения гипотезы (это тесты, используемые для измерения предположений).

- **Значение p** — это число, которое показывает, насколько сильны доказательства против предположения.

- Меньшее значение p (например, < 0,05) указывает на сильные доказательства для отклонения предположения (нулевая гипотеза), тогда как большое значение p означает, что доказательств для его отклонения недостаточно.

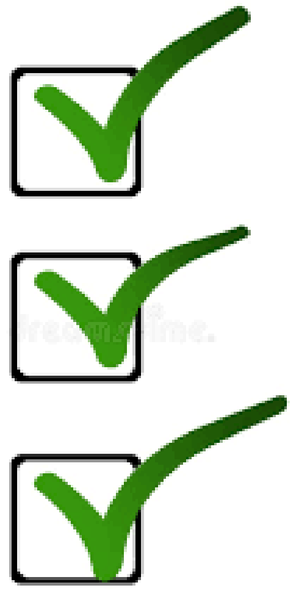
Проверка Гипотез

Пример

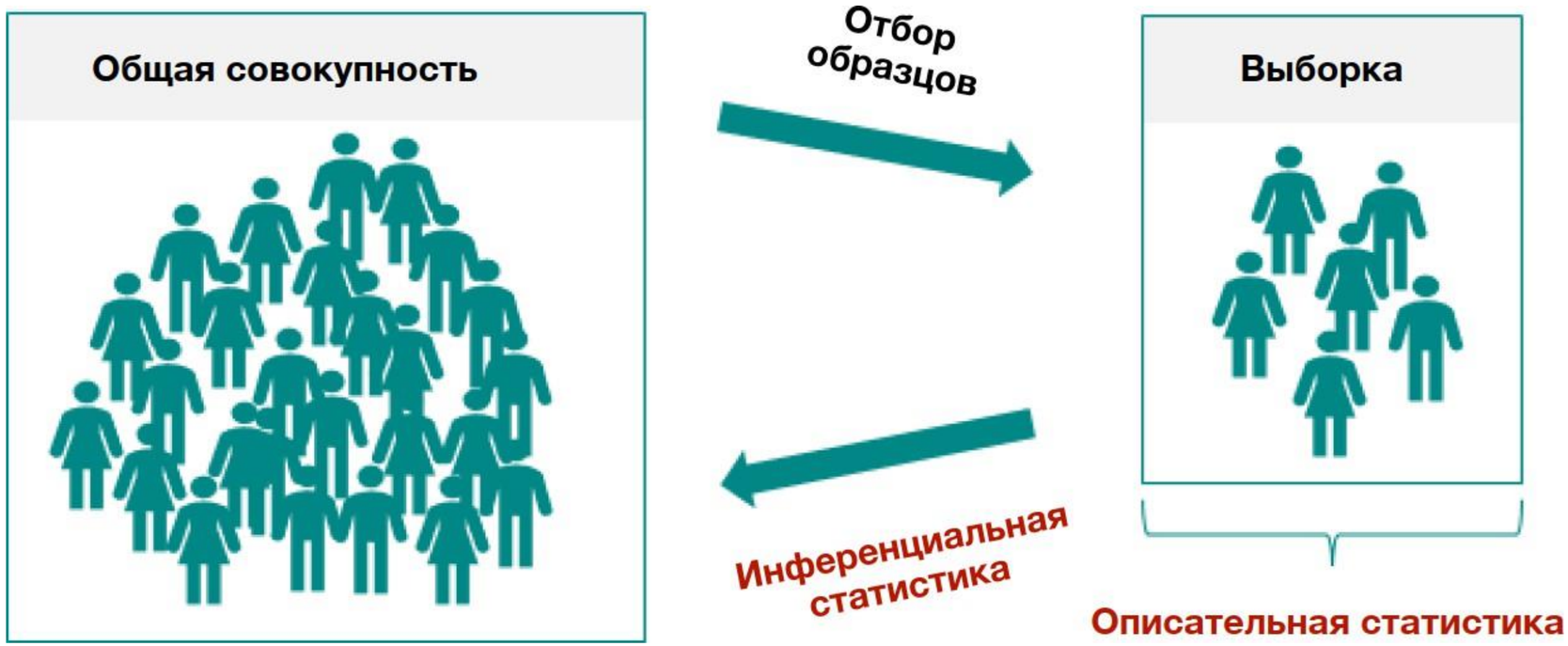
- **Исследовательский вопрос:** Планируют ли большинство зарегистрированных избирателей голосовать на следующих президентских выборах?
  - Параметр, который здесь проверяется, — это одна пропорция.
  - У нас есть одна группа: зарегистрированные избиратели.
  - «Большинство» будет больше 50%, или p>0,50.
  - Это конкретный параметр, который мы проверяем.
  - Соответствующая процедура здесь — проверка гипотезы для одной пропорции.

Стадии работы при анализе данных

- **Выборка**: вы начинаете со сбора данных из подмножества популяции, которое вас интересует для изучения. Это подмножество называется выборкой.
- **Анализ**: после сбора данных вы используете различные статистические методы. Это может включать расчет таких показателей, как средние значения, стандартные отклонения, корреляции или коэффициенты регрессии.
- **Вывод**: после анализа данных выборки вы делаете выводы или обобщения о популяции, из которой была взята выборка. Эти выводы основаны на предположении, что выборка является репрезентативной для популяции.



Стадии работы при анализе данных



Методы инференциальной статистики

- **Проверка гипотез**
- **Доверительные интервалы**

*Эти методы помогают исследователям определить, являются ли их результаты статистически значимыми и можно ли распространить их на более широкую популяцию.

Проверка гипотез

- Процесс обычно включает в себя выдвижение ***нулевой и альтернативных гипотез*** и проведение статистического теста для определения ***наличия достаточных доказательств для отклонения нулевой гипотезы в пользу альтернативной гипотезы*.**

- ***Нулевая гипотеза: Предположение по умолчанию*** *(например, «Между двумя моделями нет разницы»)*

- ***Альтернативная гипотеза: Утверждение, которое вы хотите проверить*** *(например, «Модель А работает лучше, чем модель В»).*

Проверка гипотез

Пример

- Исследователь может выдвинуть гипотезу, что средний доход людей в определенном городе превышает $50 000 в год.

  1. Он соберет выборку доходов

  2. Проведет проверку гипотезы

  3. И определит, дают ли данные достаточно доказательств, чтобы подтвердить или опровергнуть эту гипотезу.

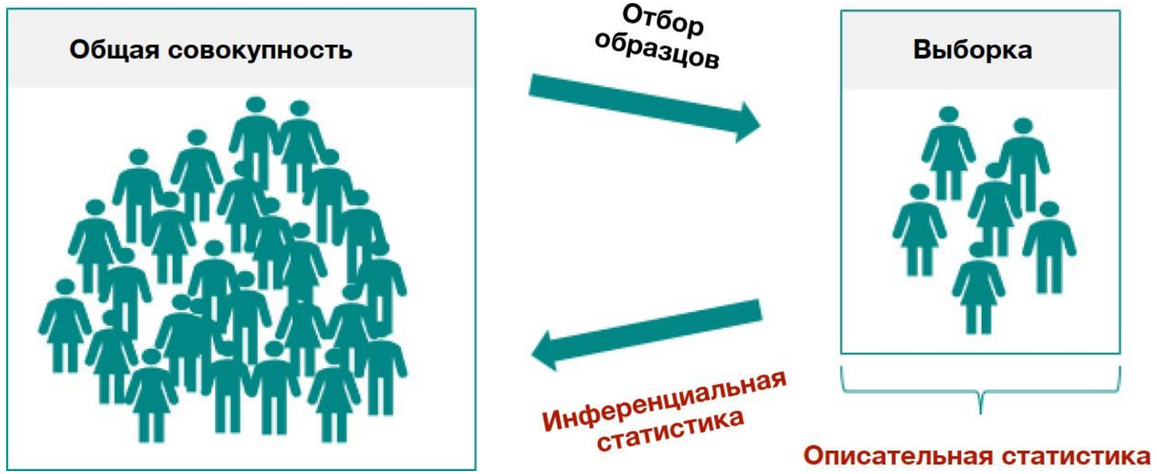
Проверка гипотезв

- Проверка гипотезы, что средний доход людей в определенном городе превышает $50 000 в год.

  1. Сбор выборки доходов

  2. Проверку гипотезы

  3. Дают ли данные достаточно доказательств, чтобы подтвердить или опровергнуть эту гипотезу.



Z-тест

- Z-тест — это статистический тест для определения того, различаются ли средние значения двух популяций, когда дисперсия популяции известна, а размер выборки большой (обычно n > 30).

- Он основан на стандартном нормальном распределении (Z-распределении).

$$Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

Где:

- $\bar{x}$ - среднее значение выборки,
- μ - среднее значение генеральной совокупности (предполагаемое),
- σ - стандартное отклонение генеральной совокупности,
- n - размер выборки.

Z-тест

- Пример: исследователь хочет определить, значительно ли отличается средний рост популяции от 165см.
- Они собирают большую выборку показателей роста с известным стандартным отклонением популяции и используют Z-тест *для сравнения среднего значения выборки со средним значением популяции.*

T-тест

- Формула для статистики t-теста похожа на Z-тест, но она использует среднеквадратическое отклонение выборки вместо среднеквадратического отклонения совокупности.
- **Пример:**

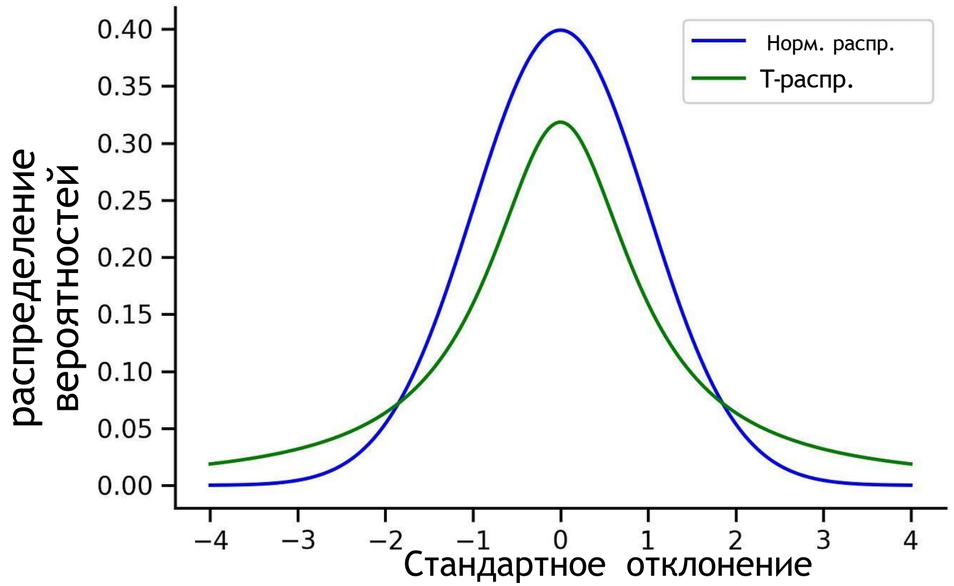
1. Исследователь хочет определить, есть ли существенная разница в результатах экзаменов между двумя группами студентов.

2. Они собирают результаты экзаменов из каждой группы и используют t-тест для сравнения средних значений.

T-тест

- T-тест используется, когда среднеквадратическое отклонение совокупности неизвестно или размер выборки мал (обычно n < 30).

- Он основан на распределении Стьюдента, которое имеет более толстые хвосты, чем стандартное нормальное распределение.



T-тест

$$t = \frac{\bar{x}_{test} - \bar{x}_{control}}{\sqrt{\frac{s^2_{test}}{n_{test}} + \frac{s^2_{control}}{n_{control}}}}$$

s_{test} — стандартное отклонение в тестовой выборке

$s_{control}$ — стандартное отклонение в контрольной выборке

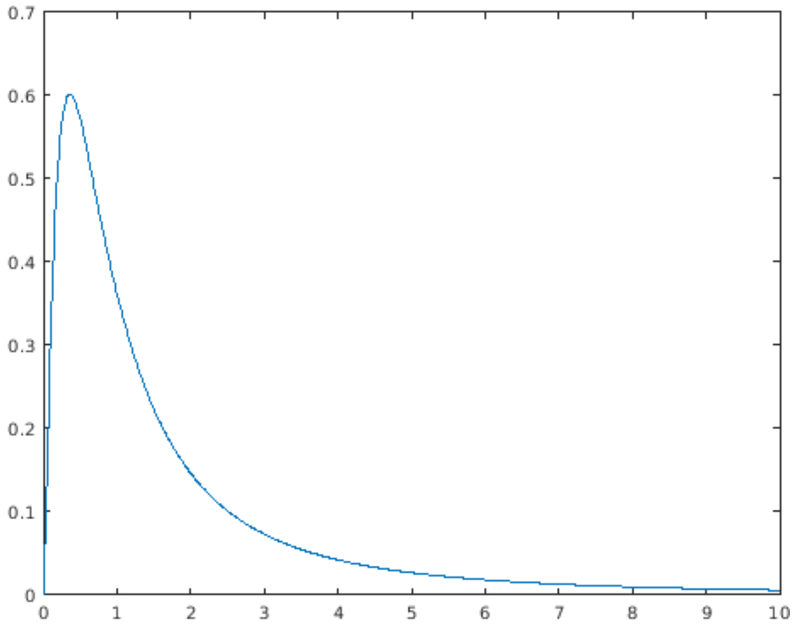
n_{test} — кол-во элементов в тестовой выборке

$n_{control}$ — кол-во элементов в контрольной выборке

F-test

- F-тест используется для сравнения дисперсий двух популяций или более двух популяций. Он обычно используется в дисперсионном анализе (ANOVA) для проверки различий между средними значениями нескольких групп.

- Статистика F-теста следует F-распределению, которое имеет положительную асимметрию и принимает только неотрицательные значения.



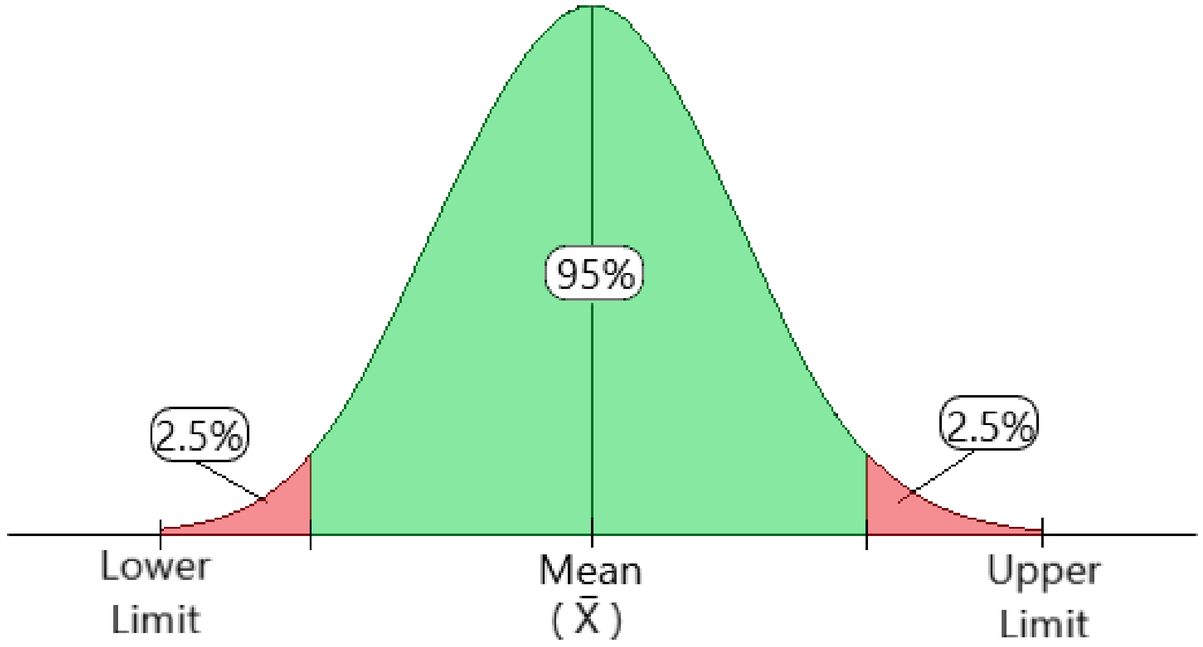
F-test

- В ANOVA F-тест сравнивает дисперсию между группами с дисперсией внутри групп. Если отношение этих дисперсий достаточно велико, это говорит о том, что средние значения групп различаются.

- Пример: исследователь хочет определить, есть ли различия в эффективности трех методов обучения для успеваемости учащихся. Они собирают данные об успеваемости учащихся, обучающихся с использованием каждого метода, и используют ANOVA, который использует F-тест, для сравнения дисперсий между группами и внутри групп.

Доверительные интервалы

- 95% доверительный интервал для среднего значения популяции указывает на то, что мы на 95% уверены в том, что истинное среднее значение популяции попадает в этот интервал.
- Доверительные интервалы используются для количественной оценки неопределенности прогнозов модели или оценок параметров.



Спасибо за внимание!